TMDB Movie Data Analysis

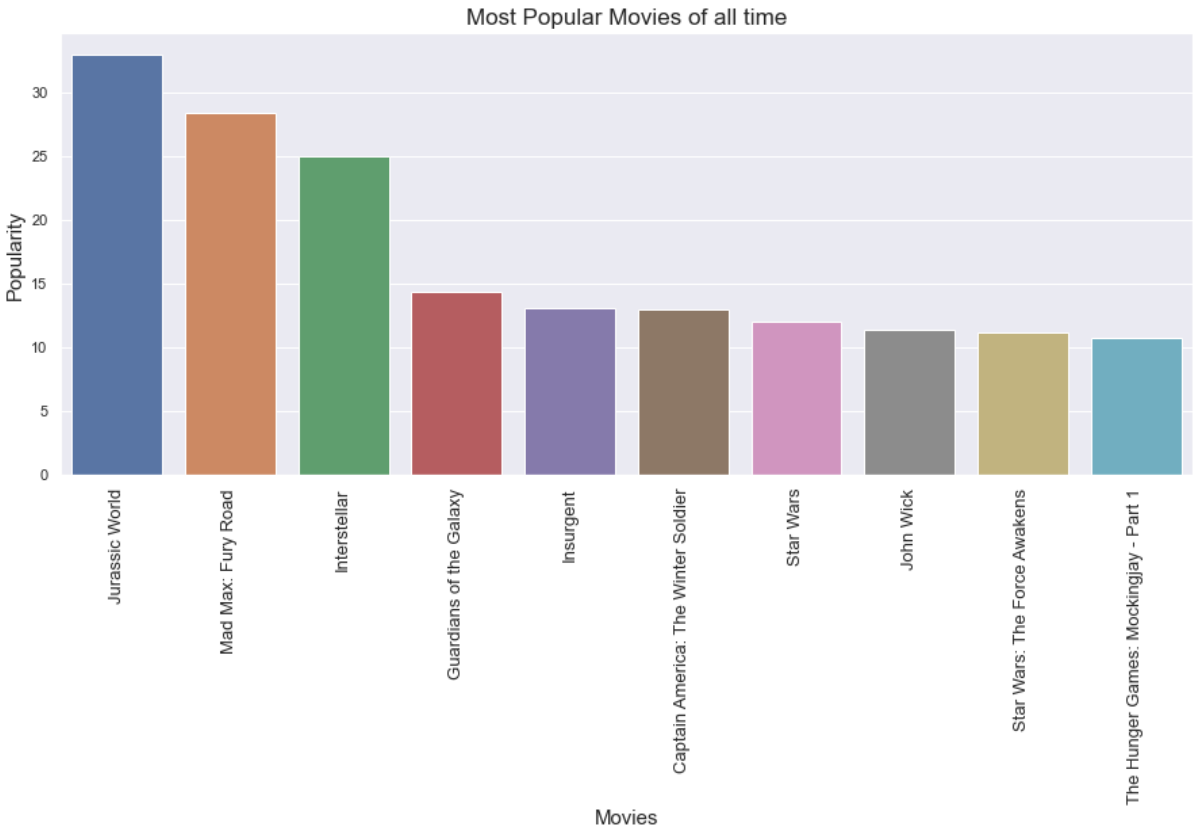
I carried out an analysis on the TMDB movie dataset. The data set contains information, about 10,000 movies obtained from The Movie Database(TMDB). The dataset consists of columns providing information on Popularity, User ratings, movie budgets, revenue, cast members, and genres. With a glance through information from the dataset I was able to pose the following questions;

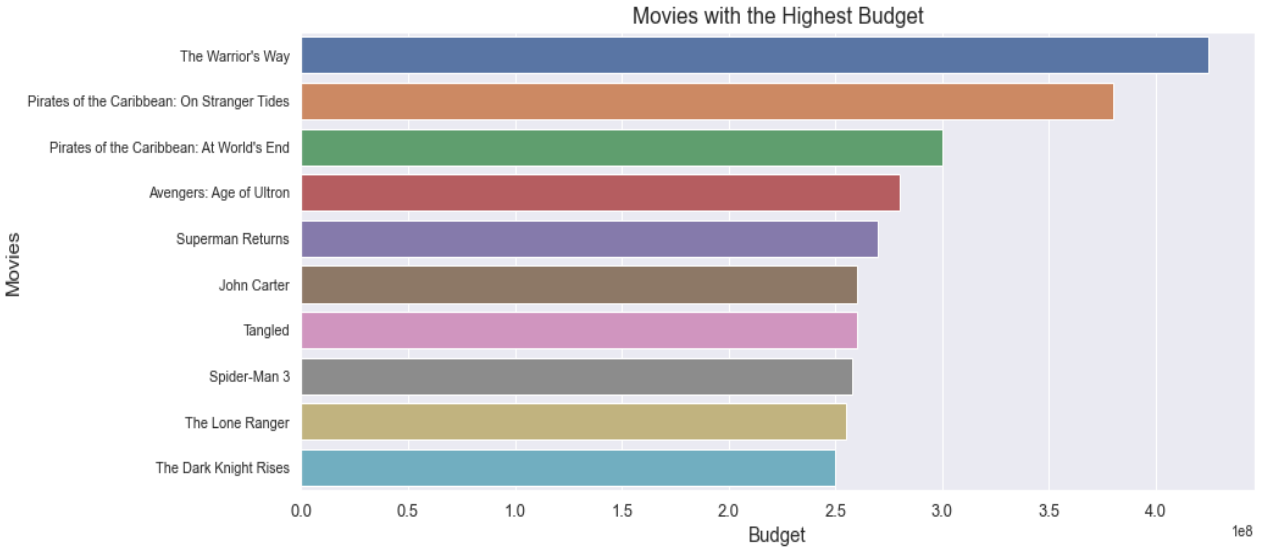
- What is the most popular movie of all time?
- What movie has the highest budget and revenue?
- Do user ratings have any effect on movie popularity?
- How do user ratings affect the revenue of a movie?
- Are movies with expensive budgets profitable?
- What is the movie with the highest Profit?
- Do User Ratings have any effect on Profit?
- Does Movie length affect the popularity of a movie?
- What is the year with the highest number of movie releases?

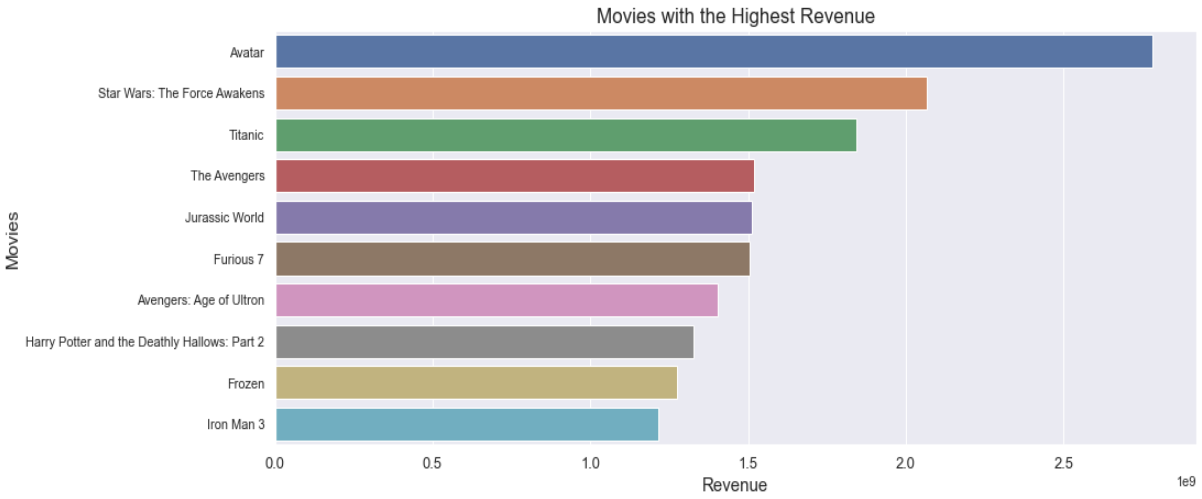
After noting these questions, I could work on the data set in tangent with the questions.
I imported, the necessary packages I needed for analysis into my jupyter notebook which were; Pandas, Numpy, Matplotlib, and Seaborn.
I used pandas, pd.read_csv() to import the data set into jupyter notebook, after downloading it.
Then I began the Data wrangling process, I cleaned the data by removing extraneous columns; imdb_id', 'homepage', 'tagline', 'keywords', 'overview', 'production_companies', 'budget_adj', 'revenue_adj.
Then I got rid of duplicates and null values. The budget, revenue, and runtime columns contained multiple zero values, so I replaced them with null values instead of dropping them, to preserve data integrity, but I deleted the zero values in the run time column as they were not as many enough to cause damage to the data integrity.
After data wrangling data were explored using visuals like bar charts, scatter plots, and line plot. Because the dataset was large it was difficult to plot every data value. So I filtered the dataset for the top 10 most popular movies, the top 10 movies with the highest budget and revenue, and the top 10 movies with the highest profit. This made it easier to plot my visuals.
To calculate profit, I subtracted the budget column from the revenue column

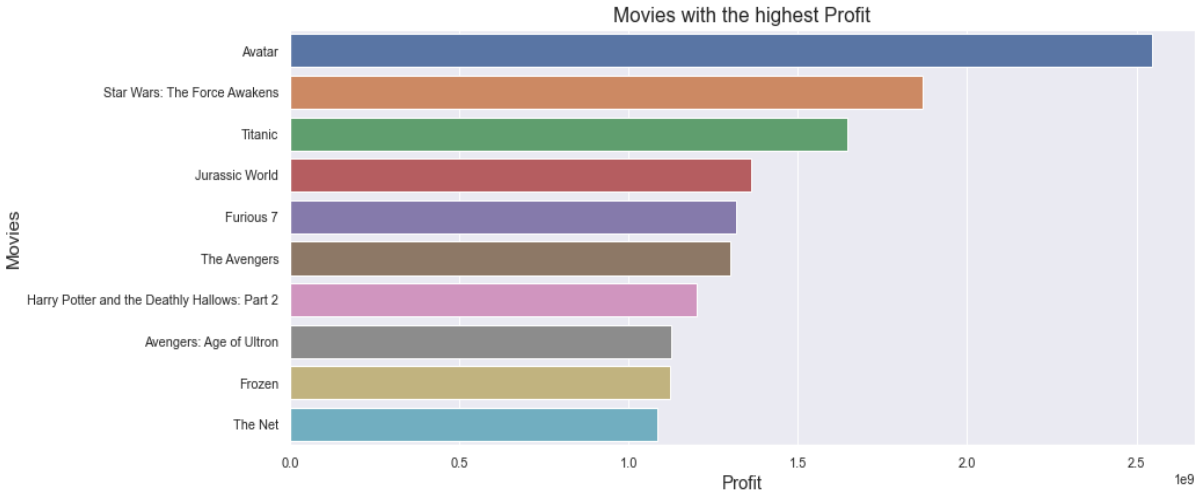
Bar charts were used to answer the following questions;

- What is the most popular movie of all time?
- What movie has the highest budget and revenue?
- What is the movie with the highest Profit?


Most Popular Movies of all time
Popularity
30
25
20
15
10
5
0
Jurassic World
Mad Max: Fury Road
Interstellar
Guardians of the Galaxy
Insurgent
Captain America: The Winter Soldier
Star Wars
John Wick
Star Wars: The Force Awakens
The Hunger Games: Mockingjay - Part 1
Movies


Movies with the Highest Budget
Movies
The Warrior's Way
Pirates of the Caribbean: On Stranger Tides
Pirates of the Caribbean: At World's End
Avengers: Age of Ultron
Superman Returns
John Carter
Tangled
Spider-Man 3
The Lone Ranger
The Dark Knight Rises
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e8
Budget

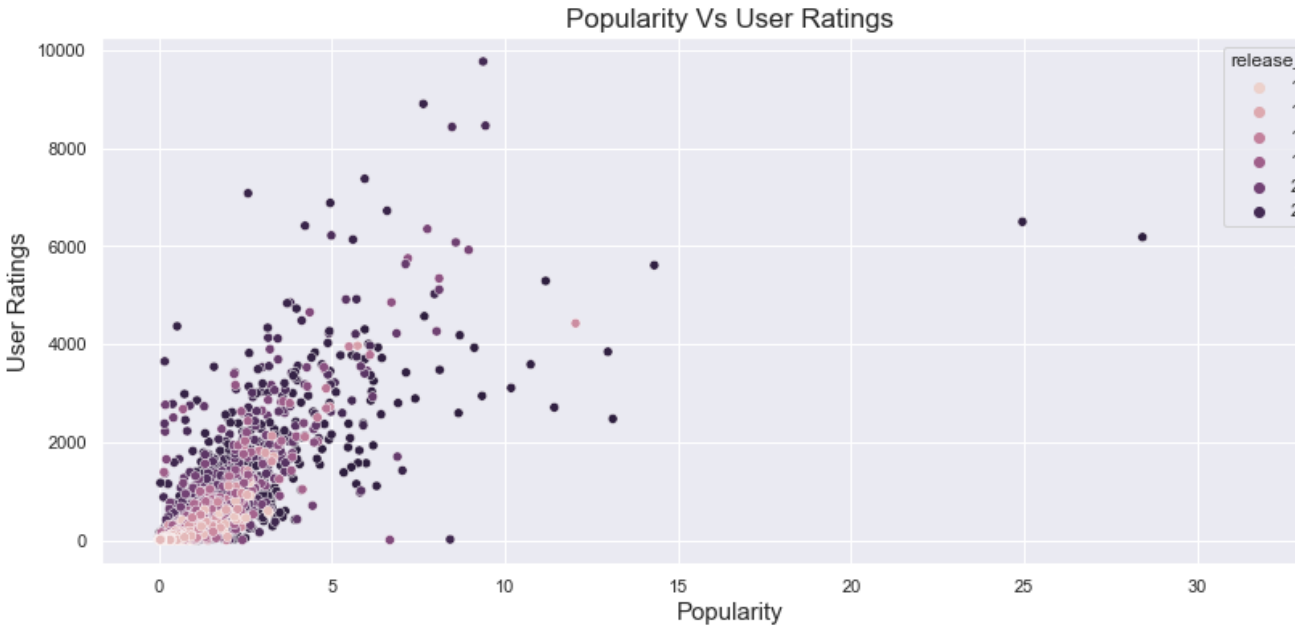


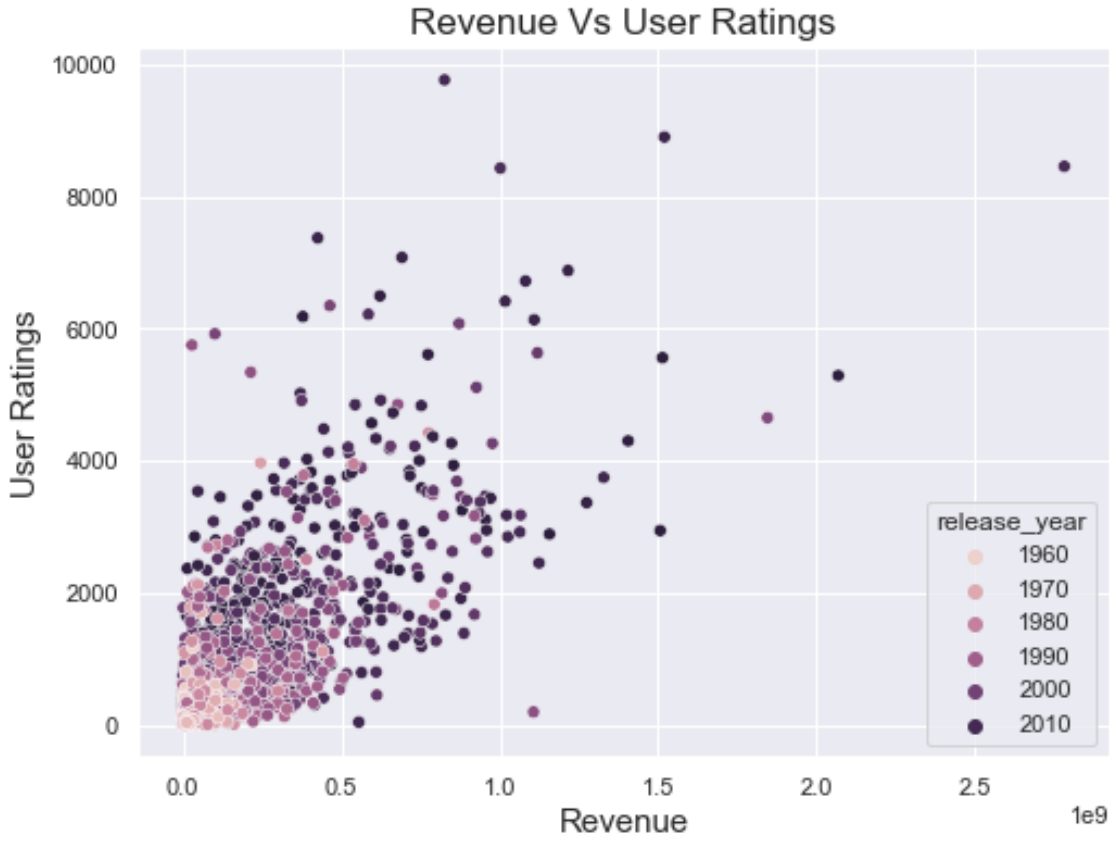


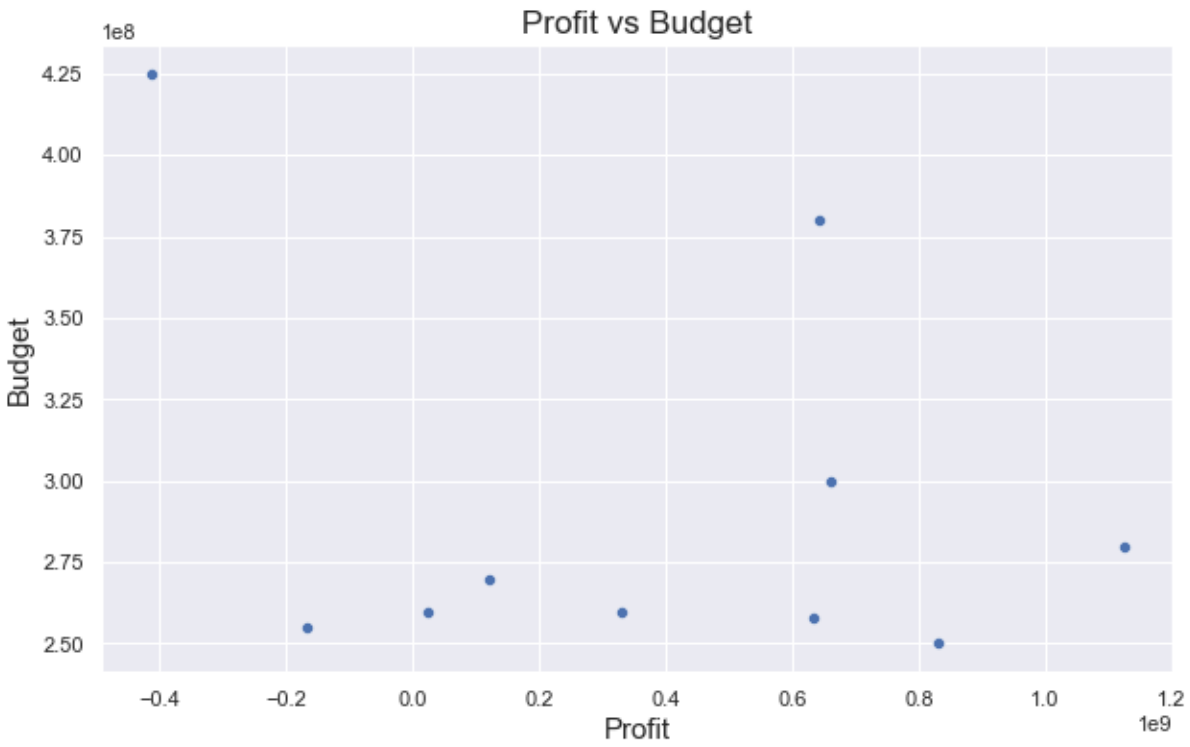
The most popular movie of all time was Jurassic Park. the movie with the highest budget was The Warrior's Way and the movie with the highest revenue and profit was Avatar.

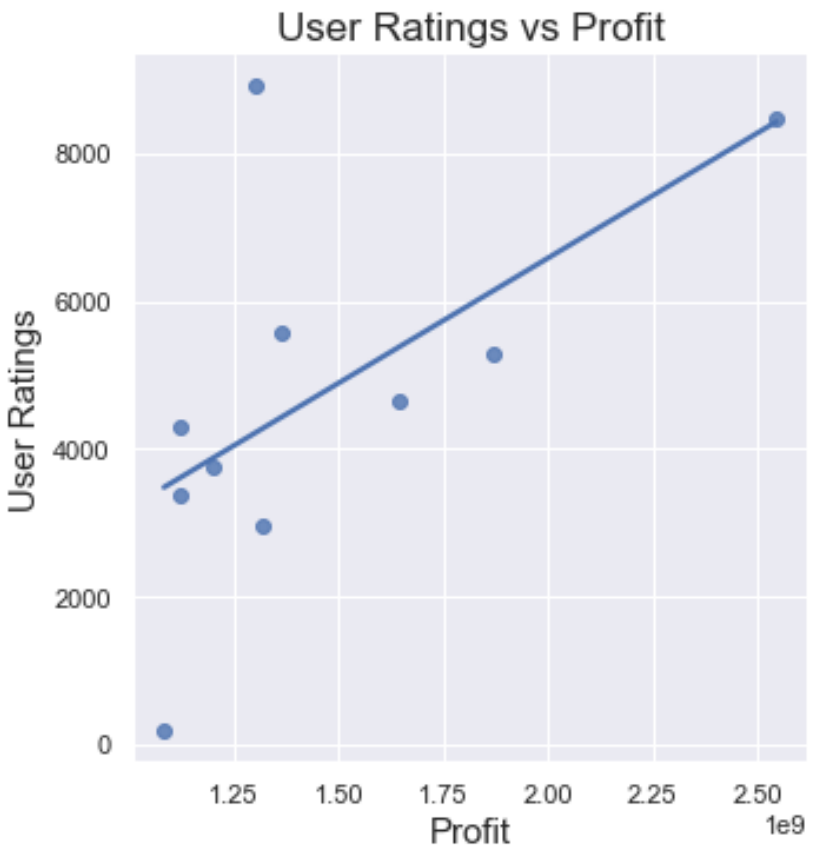
Scatter plots were used to answer the following questions;

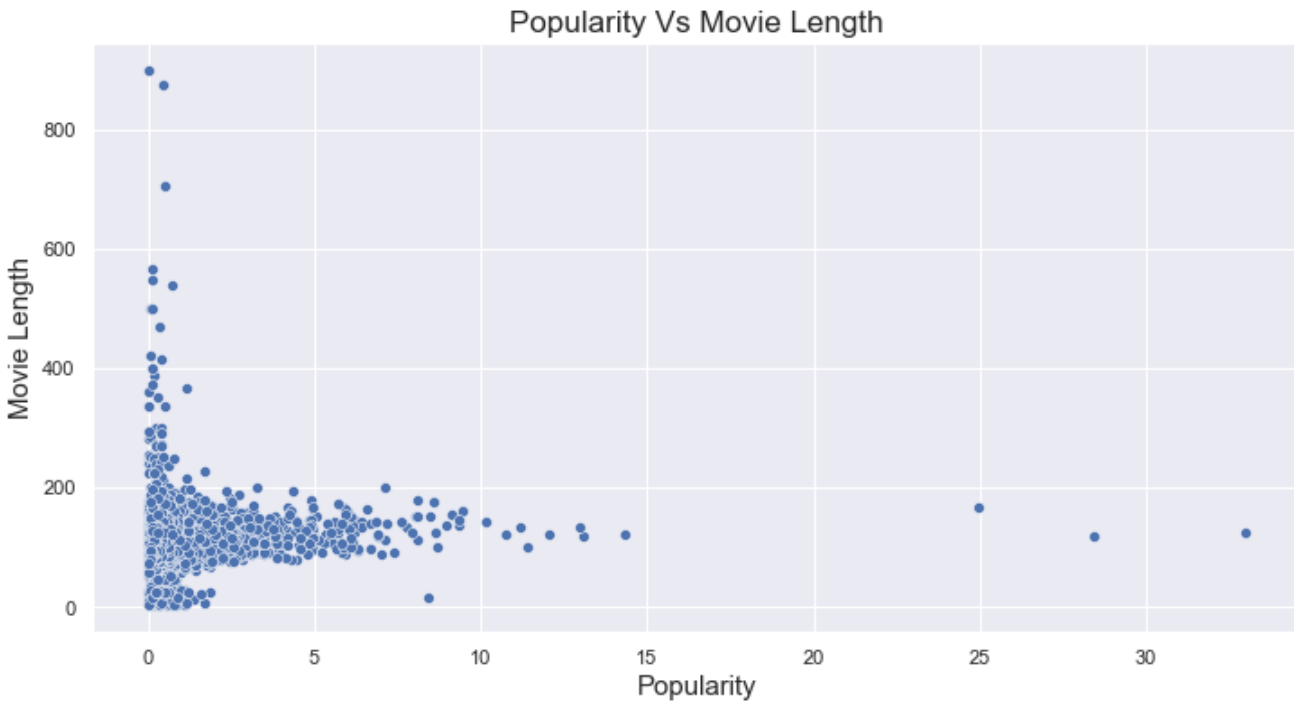
- Do user ratings have any effect on movie popularity?
- How do user ratings affect the revenue of a movie?
- Are movies with expensive budgets profitable?
- Do User Ratings have any effect on Profit?
- Does the Movie length affect the popularity of a movie?


Popularity Vs User Ratings
User Ratings
Popularity
0
2000
4000
6000
8000
10000
0
5
10
15
20
25
30


Revenue Vs User Ratings
User Ratings
Revenue
1e9
release_year
1960
1970
1980
1990
2000
2010
0
2000
4000
6000
8000
10000
0.0
0.5
1.0
1.5
2.0
2.5

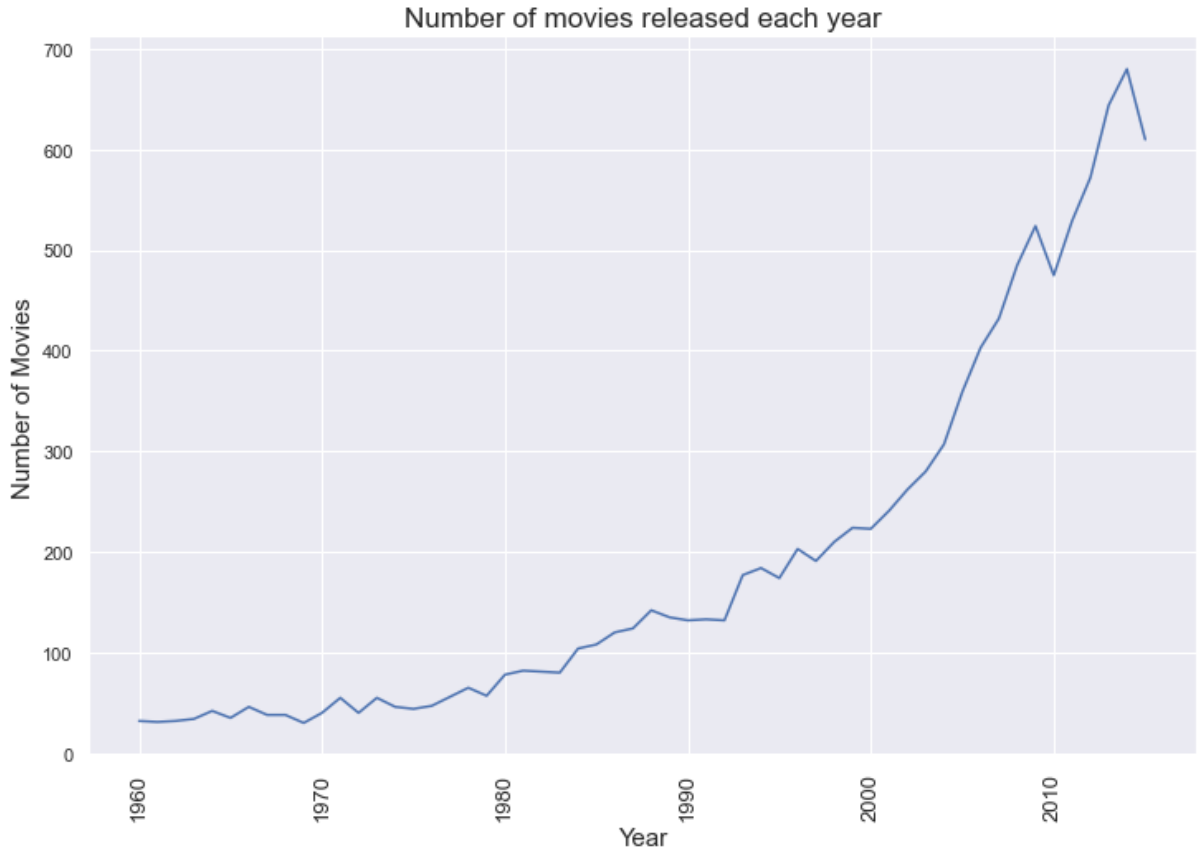

Profit vs Budget
1e8
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
Budget
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Profit
1e9


User Ratings vs Profit
8000
6000
4000
2000
0
User Ratings
1.25
1.50
1.75
2.00
2.25
2.50
Profit
1e9



There was no correlation between user ratings(vote_count) and popularity, user ratings, and revenue, runtime, and popularity. However, there was a weak correlation between user ratings and profit, not strong enough to imply causation.

Then finally line plot was used to show the year with the highest movie releases as well as the trend of movies released over the years. 2014 was the year with the highest movie releases.



Findings

The most popular movie of all time is Jurassic park, which is the fifth movie with the highest revenue. Although it was not part of the top 10 movies with the highest budget.

The movie with the highest budget: The Warrior's way, was not among the top 10 most popular movies of all time, and nor was it among the top 10 movies with the highest revenue and profit. it's easy to conclude that people didn't like it that much but that would be implying causation.

There was no correlation between User ratings and Popularity as well as User ratings and revenue. So we cannot imply causation in the context of how people's feelings affected the revenue and popularity of movies.

The movie with the highest revenue and profit is Avatar, which is ironic seeing as it wasn't among the top 10 popular movies or the movies with the highest budget, which proves that there is no correlation in this case

between profit and budget. Although there is a weak correlation between user ratings and profit, we cannot still imply that Avatar had the highest profits because it had high user ratings.

From our top 10 samples, we also observed that not all movies with expensive budgets are profitable, honorable mention, The Warrior's Way

Furthermore based on our visualization, we conclude that the length of a movie, that is runtime, has no effect on the popularity of the movie

Finally, we see that there is an increase in the trend of the number of movies released each year. This shows that over the years, the movie industry has thrived in an impressive way, proving that there is a demand for more each year.

Limitations

The Genres and Cast columns contained pipe characters(|), so it was difficult to explore the area of that dataset.

Data Integrity: There were so many zero values in the budget and revenue columns, which had to be replaced with null values.

There was no clarity on the units of the budget and revenue columns.

I did not consider the vote_average column so the rating score may be biased.

I ignored the budget_adj andt revenue_adj columns

I also did not really explore the properties that contributed to a movie's success